Introduction

Boston University CS 506 - Lance Galletti

Data Representation

How we represent data is linked to what information we are able to retrieve from it.

Data Representation - Records

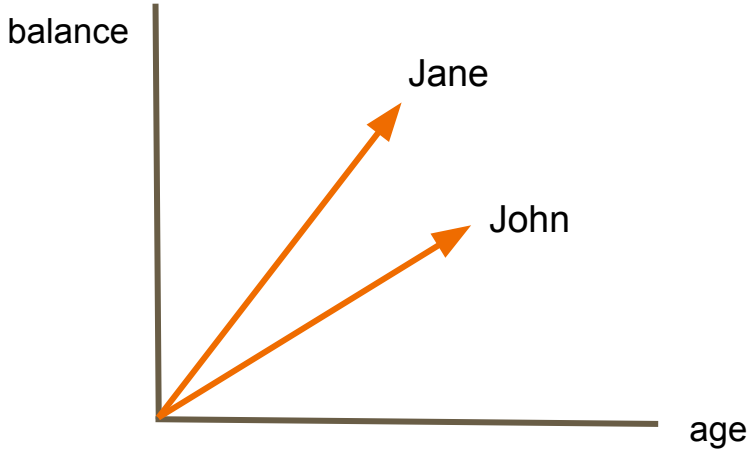
m-dimensional points / vectors

Example: (name, age, balance) -> (“John”, 20, 100)

Data Representation - Records

m-dimensional points / vectors

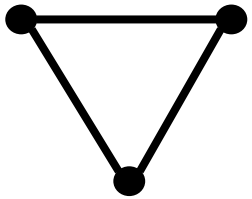
Example: (name, age, balance) -> (“John”, 20, 100)



Data Representation - Graphs

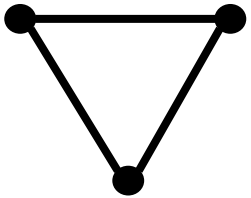
Nodes connected by edges

Example:



Adjacency Matrix

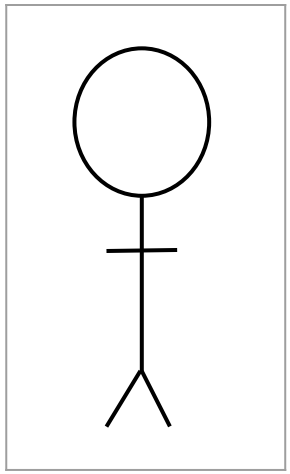
$$\begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}$$



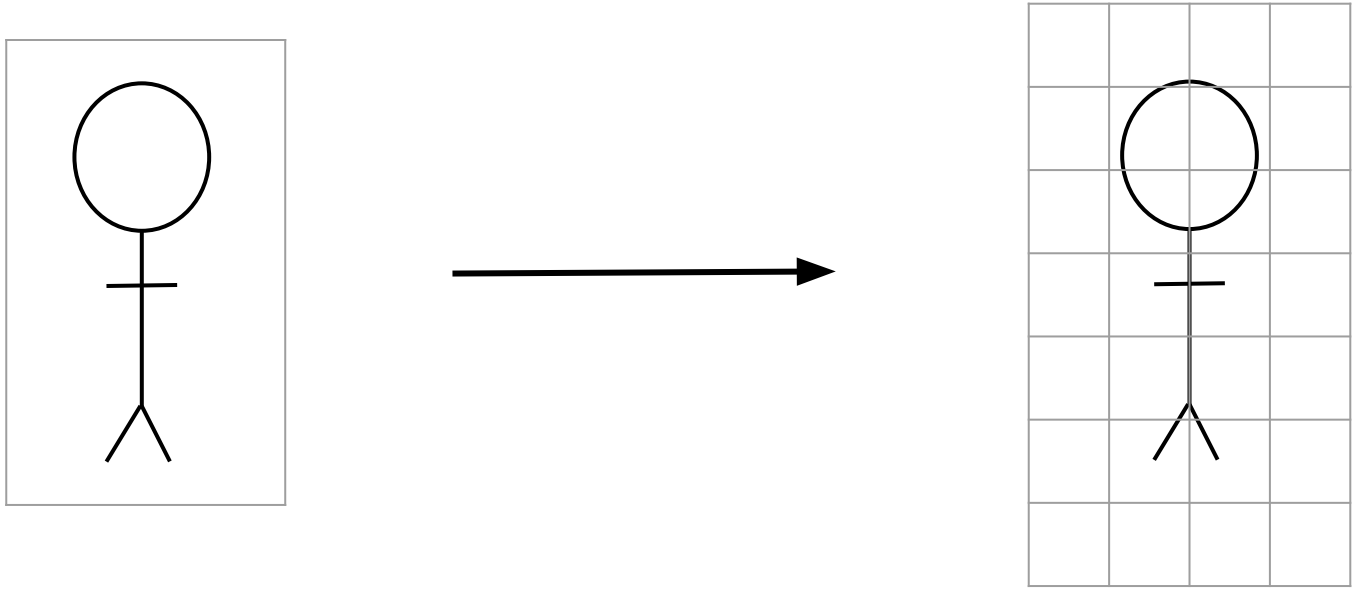
Adjacency List

1 : {2, 3}
2 : {1, 3}
3 : {1, 2}

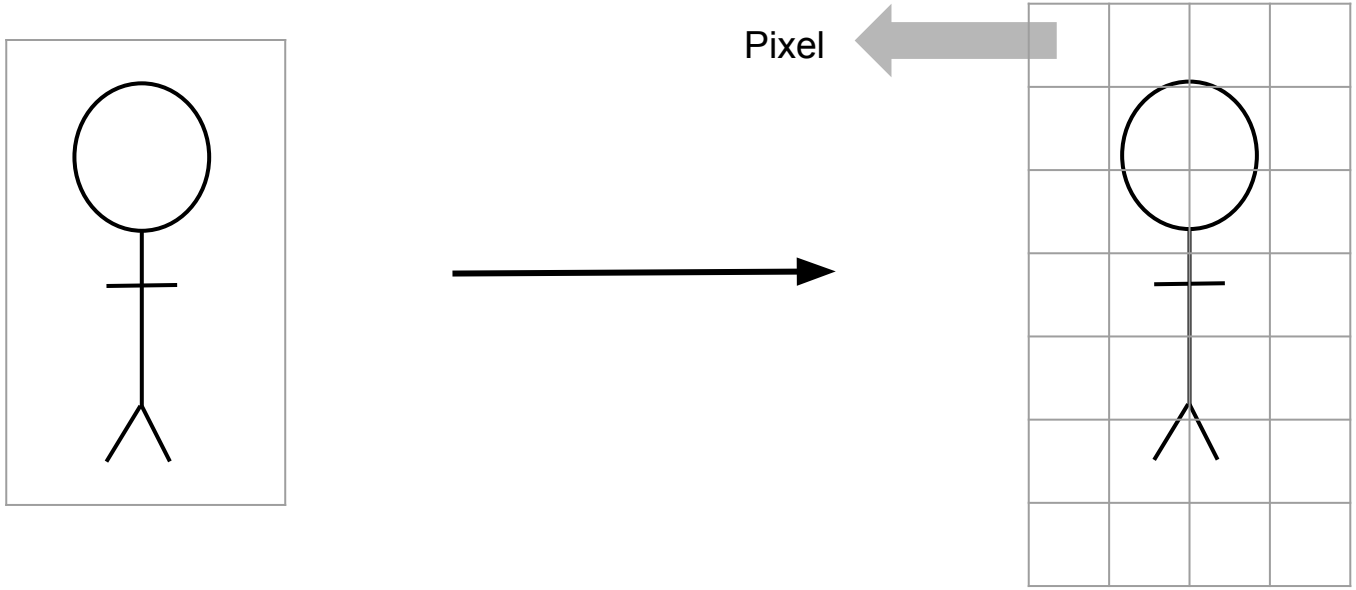
Data Representation - Images



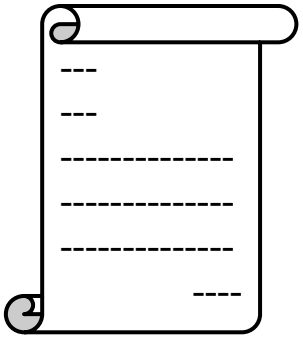
Data Representation - Images



Data Representation - Images



Data Representation - Text



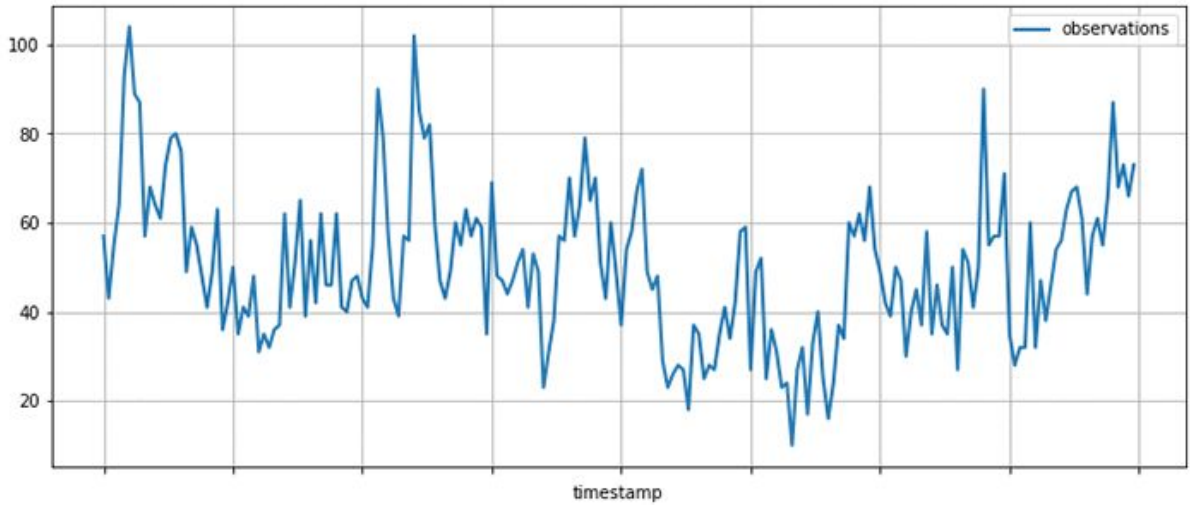
List of words

Data Representation - Strings

DNA seq (A T G C C G T A …) -> list of characters

Data Representation - Time Series

List of data at specific intervals of time



Types of Learning

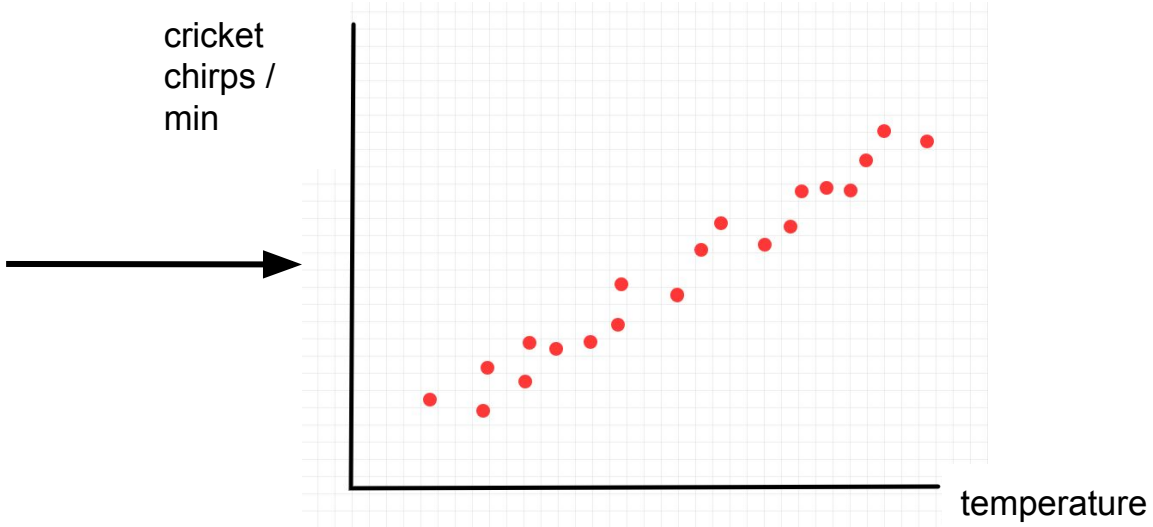
- Supervised Learning
- Unsupervised Learning

Supervised Learning

cricket chirps / min	temperature
10	40
5	37
17	53
55	103
40	78

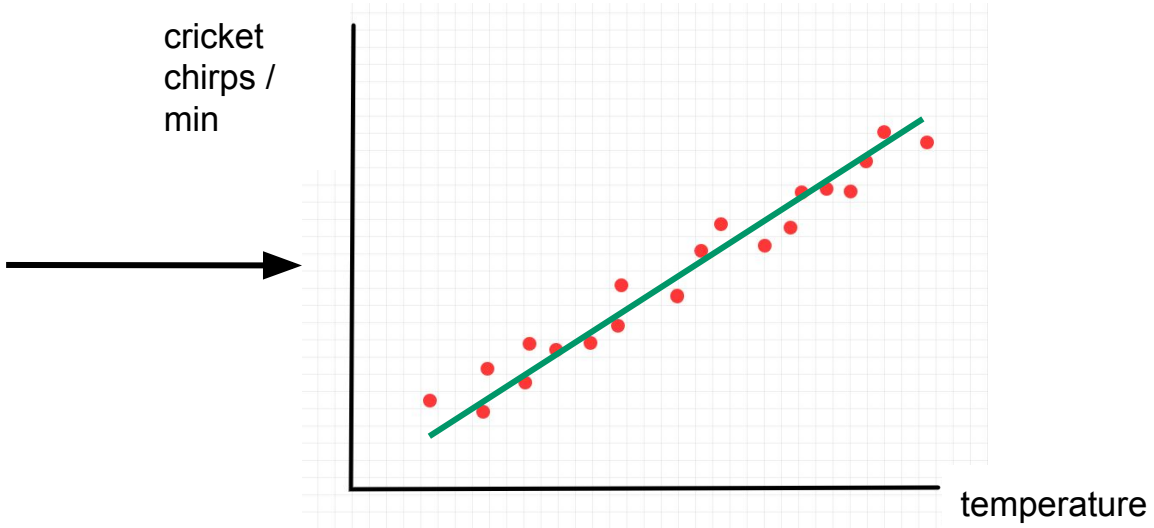
Supervised Learning

cricket chirps / min	temperature
10	40
5	37
17	53
55	103
40	78



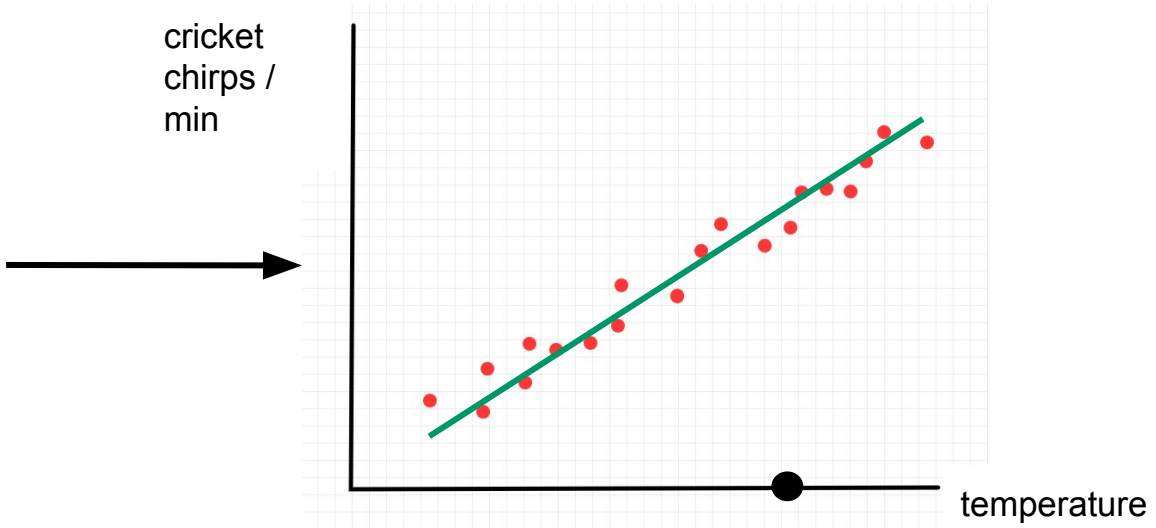
Supervised Learning

cricket chirps / min	temperature
10	40
5	37
17	53
55	103
40	78



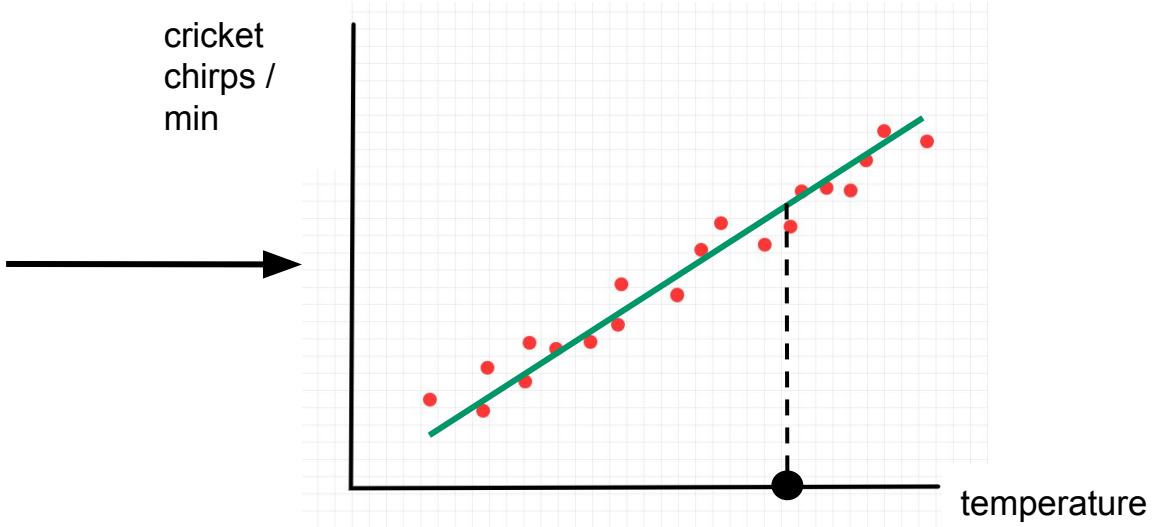
Supervised Learning

cricket chirps / min	temperature
10	40
5	37
17	53
55	103
40	78



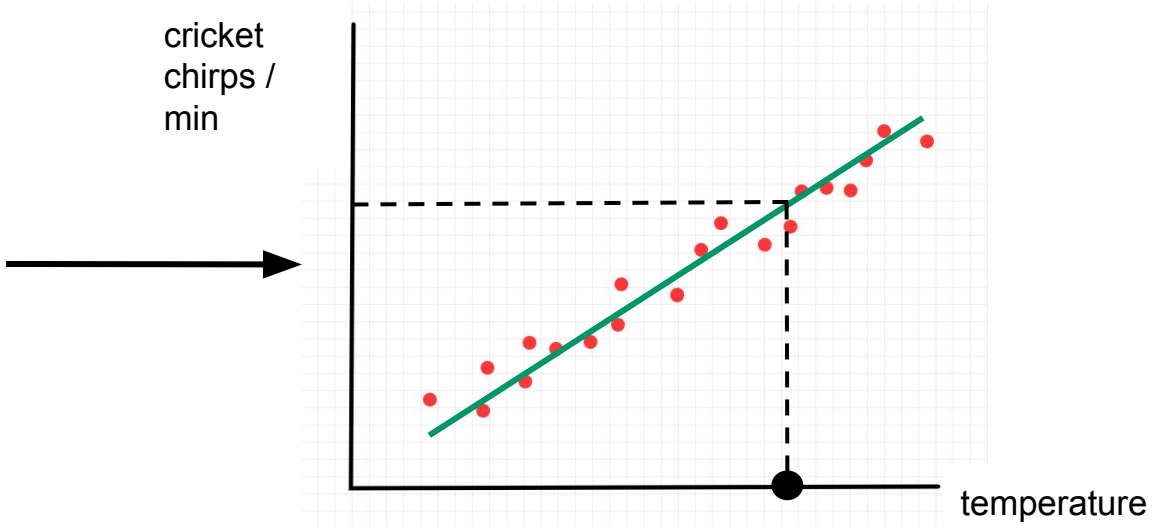
Supervised Learning

cricket chirps / min	temperature
10	40
5	37
17	53
55	103
40	78



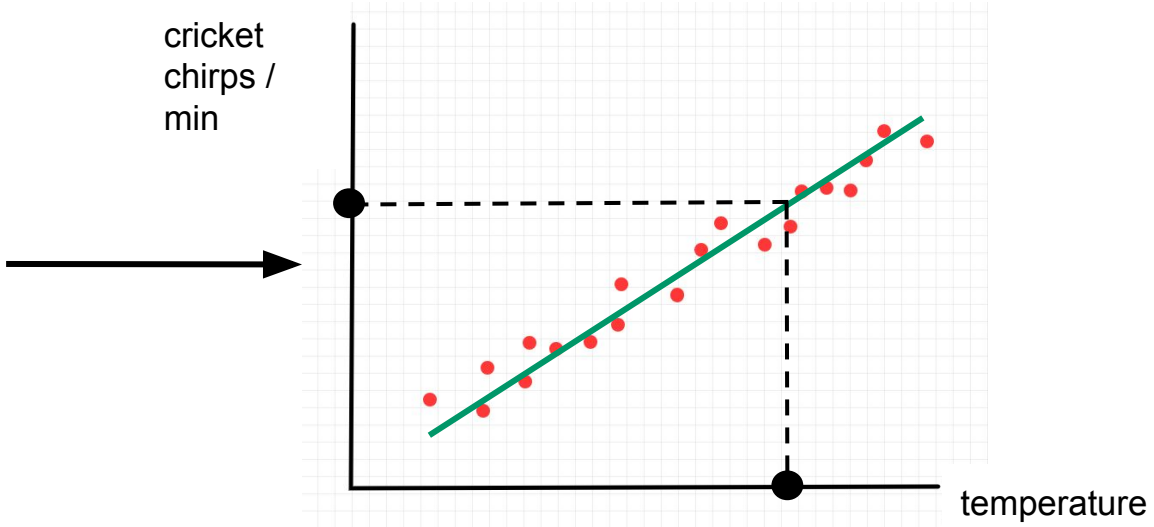
Supervised Learning

cricket chirps / min	temperature
10	40
5	37
17	53
55	103
40	78



Supervised Learning

cricket chirps / min	temperature
10	40
5	37
17	53
55	103
40	78



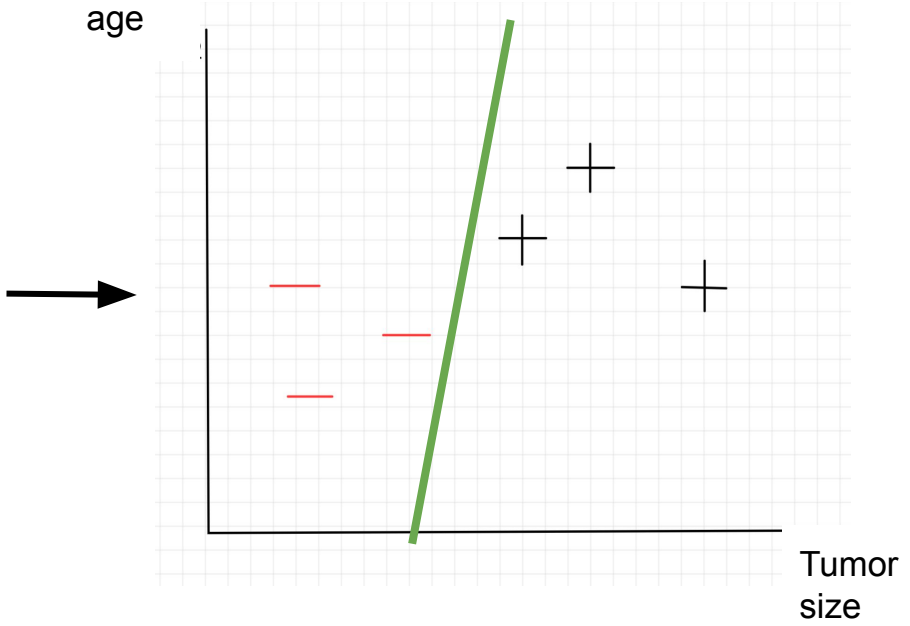
This type of supervised learning is referred to as regression

Supervised Learning

age	tumor size	malignant
20	12	0
22	15	1
47	20	1
59	2	1

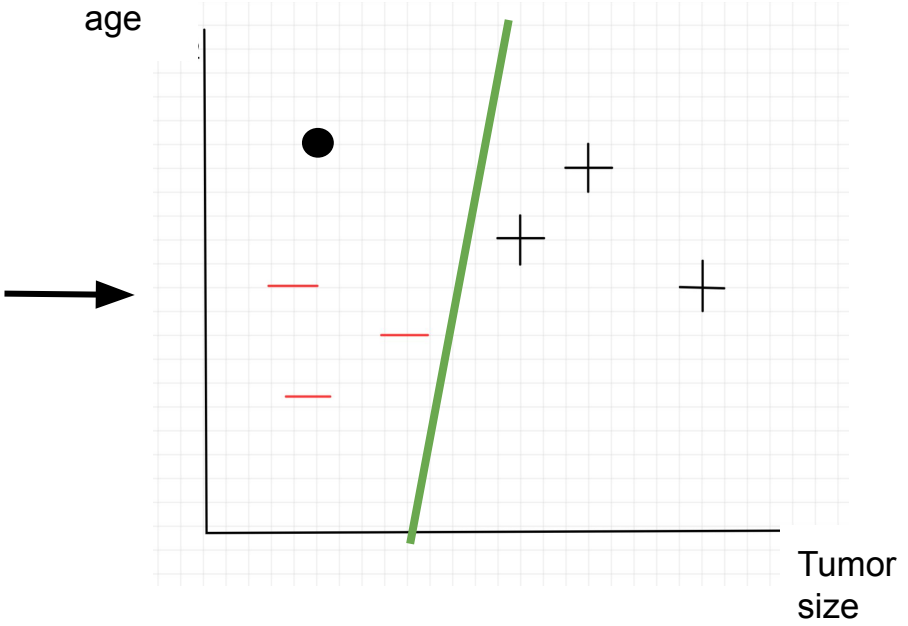
Supervised Learning

age	tumor size	malignant
20	12	0
22	15	1
47	20	1
59	2	1



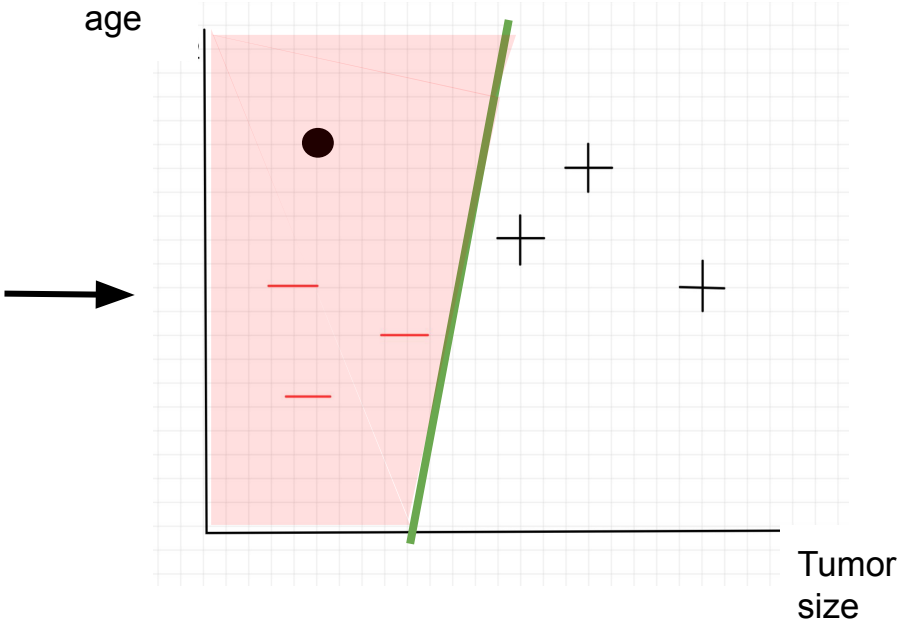
Supervised Learning

age	tumor size	malignant
20	12	0
22	15	1
47	20	1
59	2	1



Supervised Learning

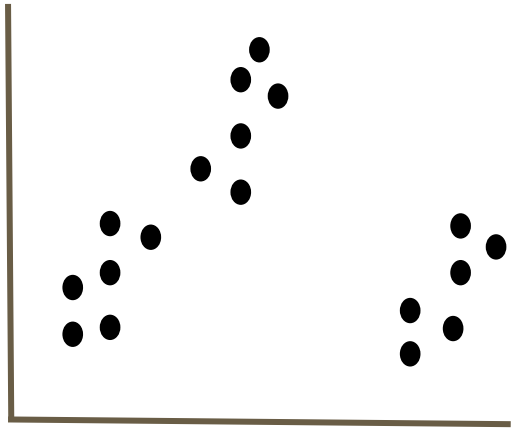
age	tumor size	malignant
20	12	0
22	15	1
47	20	1
59	2	1



This type of supervised learning is referred to as classification

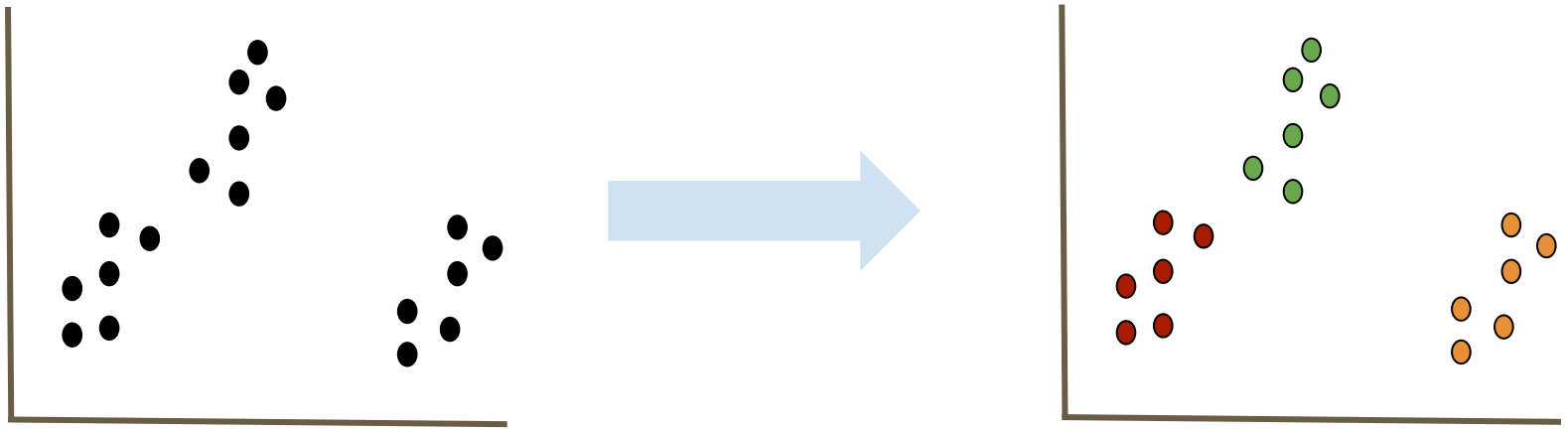
Unsupervised Learning

Goal: Find interesting structure in the data



Unsupervised Learning

Goal: Find interesting structure in the data



This type of unsupervised learning is referred to as clustering

Unsupervised Learning

Dataset: Collection of Articles

Question: Are these articles covering the same topics?